“弄假成真”:基于对抗学习的数据增广方法

刘勇,曾仙芳

浙江大学智能系统与控制研究所,杭州 310027

摘要　近年来,深度学习在计算机视觉领域取得了巨大的突破,其背后是利用大量标签数据对深度网络进行监督训练,而标注大规模数据集非常昂贵且十分耗时。针对大规模数据集标注问题,苹果公司的Shrivastava团队希望借助现有的计算机仿真技术以及对抗训练的方法,实现仿真图像的无监督学习,从而避免昂贵的图像标注过程。该团队在对抗网络的基础上提出3个创新点:(1)自正则项;(2)局部对抗损失;(3)使用历史生成图片更新判别器,使得生成真实化图片的同时保留输入图像特征。实验结果展示该方法可以生成高度真实化的图片。研究者通过训练凝视估计模型、手部姿态估计模型定量分析生成图片的效果,分析结果表明,使用生成图片训练的模型,在MPIIGaze数据集上测试效果有很大的提升,达到了当时最好的效果。不过,研究者并未在包含多个物体的复杂场景下进行实验,文中提出的方法在复杂场景下的应用还存在局限性。

关键词　对抗训练;无监督学习;仿真图片;深度学习

近年来,深度学习在计算机视觉许多领域取得突破性进展。2017年7月23日,计算机视觉领域顶级会议——计算机视觉与模式识别国际会议(IEEE Conference on Computer Vision and Pattern Recognition, CVPR)公布了2017年会议最佳论文,其中一篇是苹果公司Shrivastava团队[1]的论文《Learning from simulated and unsupervised images through adversarial training》,这篇文章引起了领域内学者的广泛关注。该文尝试解决训练大型深度网络需要大量标签数据的问题。研究表明,通过对抗训练[2]的方式提高仿真图片的真实性,从而深度网络可以从无标签仿真图片中学习知识,提高真实场景下的识别能力。

1　仿真图像无监督学习

深度学习在计算机视觉领域取得巨大成功,它利用大量标签数据对深度网络进行监督训练,而标注大规模数据集非常昂贵和耗时。研究者希望借助现有计算机仿真技术帮助解决此问题,现有的计算机仿真技术利用计算机图形学原理[3],可以模拟物体的整体结构特征,而对物体局部细节的模拟不够逼真。直接利用仿真图片训练深度网络,得到视觉模型会过拟和不逼真的仿真图像细节,降低了在真实场景下的识别能力。为了能更好地从这些仿真图片中学到知识,研究者利用真实图片对抗训练[4]提高仿真图片的逼真性。

收稿日期:2018-07-15;修回日期:2018-08-15

作者简介:刘勇,教授,研究方向为机器学习、机器人视觉,电子信箱:Yongliu@iipc.zju.edu.cn;曾仙芳(共同第一作者),博士研究生,研究方向为计算机视觉,电子信箱:zzlongjuanfeng@zju.edu.cn

引用格式:刘勇,曾仙芳. “弄假成真”:基于对抗学习的数据增广方法[J]. 科技导报, 2018, 36(17): 19-22; doi: 10.3981/j.issn.1000-7857.2018.17.003

研究者在论文中设计了一个提高仿真图片真实性的结构——SimGAN,其核心思想是采用对抗训练,利用深度网络提高仿真图片的真实性(图1)。SimGAN用仿真器(simulator)生成仿真图片,仿真图片作为输入图片输入精炼网络(refiner),精炼网络是一种特殊设计的深度网络,其通过对抗训练可以改善输入图片的局部细节,输出更为逼真的仿真图片(图2)。

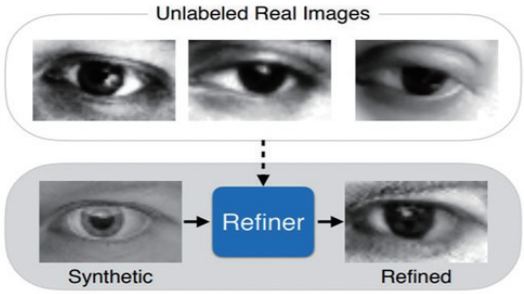


图1 SimGAN:生成高真实度的仿真图片

Fig. 1 SimGAN to generate real simulated image

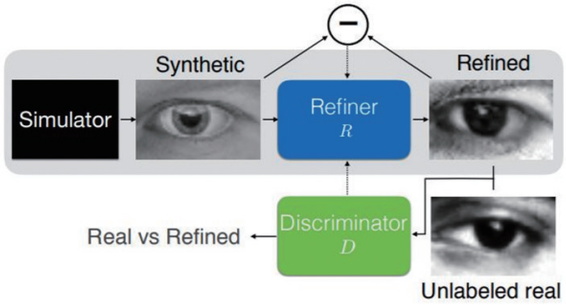


图2 SimGAN流程

Fig. 2 SimGAN pipeline

精炼网络的训练有2个损失项来约束:(1) 对抗损失项,通过引入额外的判别网络(discriminator)对比真实图片和仿真图片,对生成图片的真实性进行评判,约束精炼网络的输出图片具有更真实的细节;(2) 重构损失项,通过对比精炼网络的输出图片和输入图片,计算精炼网络的重构误差,约束精炼网络的输出图片,保持原有图片的结构信息。

以MPIIGaze[5]数据集的眼球图片为例,训练眼球图片的精炼网络过程中,对抗损失项使精炼网络的输出图片更真实化,例如,在眼角的位置呈现阴影,眼球的纹理更细致等,而重构损失项就是确保精炼网络的输出图片还是眼球图片,例如,存在瞳孔、眉毛等结构。对于NYU数据集中的手部姿态图片来说,由于传感器的限制,真实的图片数据存在空洞和边缘不平整的现象,而计算机仿真图片边缘均是十分光滑,和真实图片存在差异。论文中实验结果如图3、图4所示。

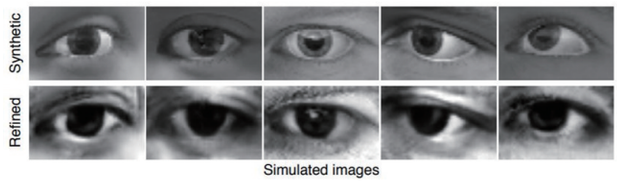


图3 眼球图片精炼网络的输出和仿真图片的比较

Fig. 3 Refinement results on UnityEye dataset

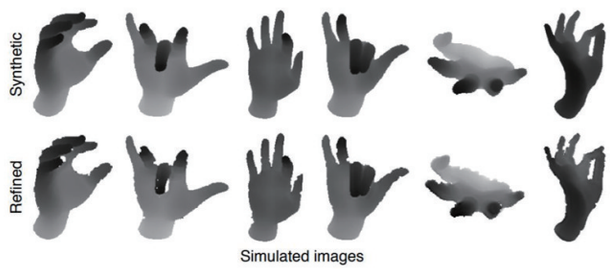


图4 手势深度图片精炼网络的输出和仿真图片的比较

Fig. 4 Refinement results on NYU hand pose dataset

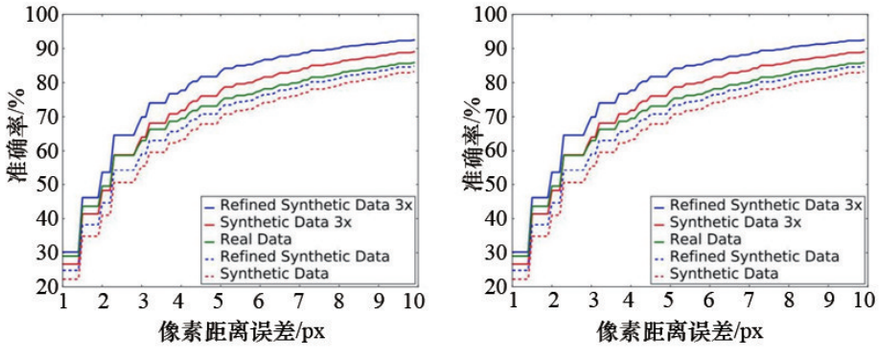
从视觉感知上来看,精炼网络确实提高了输入图片的真实感,在眼球精炼示例中,仿真图片中的瞳孔与虹膜有比较清楚的轮廓边界,而真实场景下的眼球图片由于像素、光照等原因,不会呈现如此清楚的轮廓边界。在手势深度图精炼示例中,仿真图片中手掌的边缘总是平整而光滑的,但实际通过传感器采集的图片,由于传感器数据存在噪声的原因,在手掌的边缘会存在不光滑和空洞的现象。

研究者还设计了2个结果评估实验,定量分析精炼网络输出后的图片对训练深度神经网络的作用。眼球图片精炼的结果分析实验有4组不同数据,分别是:第1组,仿真图片数据;第2组,仿真图片数据,数据量为第1组的4倍;第3组,精炼网络输出的图片数据;第4组,精炼网络输出的图片数据,数据量为第3组的4倍。

手势深度图片精炼实验采用NYU[6]数据集,手势深度图片精炼的结果分析实验有5组不同的数据,分别是:第1组,仿真图片数据;第2组,仿真图片数据,数据量是第1组数据的3倍;第3组,精炼网络输出的图片数据;第4组,精炼网络输出的图片数据,数据量为第3组的3倍;第5组,真实手势图片。

实验结果的横轴是误差的阈值,纵轴是深度网络

的准确率,从实验结果可以看出,用精炼网络的输出图片训练深度网络,其测试结果要优于用原始仿真图片训练深度网络的测试结果;此外,由于仿真图片的获取是十分容易的,研究者还测试了大量数据对训练深度网络的增益,在手势深度图片实验中,采用3倍于真实图片的精炼后仿真图片训练深度网络,其效果优于直接用真实图片训练深度网络(图5)。



(a) 眼球实验结果　(b) 手势深度图实验结果

图5　眼球及手势深度图的实验结果

Fig. 5　Eye gaze estimation results and hand posture estimation results

2　结果分析与评价

实验结果表明,利用现有的计算机图形学技术,可以合成大规模多样化的仿真图像,但是仿真图像的逼真性不够,直接用来训练深度网络,得到的结果会过拟合仿真图像的细节,在真实场景下应用效果并不理想。通过对抗训练的方式,提高仿真图片的逼真性,使得仿真图片逼近真实图片的水平,用精炼后的图片来训练深度网络,可以得到比较好的效果。甚至由于仿真图片的易获取性,采用大量精炼后的仿真图片训练深度网络,效果优于采用真实图片训练的深度网络,例如,论文中手势深度图片的定量分析实验。

论文提出的方法对利用仿真图片训练深度网络提供了一定启发,利用计算机可以高效地产生丰富多样的仿真图片,在产生图片的同时,仿真图片自带精确的标签,将仿真图片和对抗训练相结合是辅助深度网络训练的重要手段,也将是今后研究的热点。

论文提出的方法也存在一定局限性,对抗训练的提出受博弈论中零和博弈的启发,约束生成网络和判别网络不断博弈,使生成网络的输出逼进真实图片的分布,由于同时交替训练2个深度网络,对抗训练的训练过程不太稳定,而且容易出现模型崩溃的问题(模式崩溃是指生成的图片塌缩至某几个样本上)。此外,利用计算机生成仿真图片来辅助训练深度网络,需要在仿真系统里对物体建模,使其在某些场景下不太适用,例如,城市范围内的建筑物识别,如果采用仿真图片来辅助训练,需要对城市内各个建筑建立仿真模型,其工作量是巨大的,此方法便不太适用。

3　展望

《Learning from simulated and unsupervised images through adversarial training》一文利用对抗训练的方法从无标记的仿真图片中学习知识,提出的精炼网络结合判别器提高了仿真图片的逼真性,同时保留了物体的结构信息。将仿真图片和对抗训练相结合辅助深度网络训练是今后的研究热点,下一步研究重点可能在以下3个方面。

1) 探寻更有效的生成模型,使生成的图片更加多样化,训练的过程更加稳定,生成的图片真实度更高。

2) 尝试生成更为复杂的场景,同一场景中包含多个物体。

3) 尝试仿真系统和对抗训练更有效的结合方式,例如,用仿真物体的视频代替图片来辅助深度网络的训练。

参考文献(References)

[1] Shrivastava A, Pfister T, Tuzel O, et al. Learning from simulated and unsupervised images through adversarial training[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway NJ: IEEE, 2017: 2242-2251.

[2] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//International Conference on Neural Information Processing Systems. Boston: MIT Press, 2014: 2672-2680.

[3] Wood E, Baltrušaitis T, Morency L, et al. Learning an appearance-based gaze estimator from one million synthesised images [C]//Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications. New York: ACM, 2016: 131-138.

[4] Wood E, Morency L P, Robinson P, et al. Learning an appearance-based gaze estimator from one million synthesised images [C]//Biennial ACM Symposium on Eye Tracking Research & Applications. New York: ACM, 2016: 131-138.

[5] Zhang X, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway NJ: IEEE, 2015: 2201-2231.

[6] Tompson J, Stein M, Lecun Y, et al. Real-time continuous pose recovery of human hands using convolutional networks[J]. ACM Transactions on Graphics, 2014, 33(5): 1-10.

From simulation to real, adversarial learning based data augmention method

LIU Yong, ZENG Xianfang

Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China

Abstract Deep learning has recently made a huge breakthrough in the field of computer vision. What makes it succeed is using a large amount of labeled data for supervised learning with deep neural networks. However, labeling a large-scale dataset is very expensive and time-consuming. To solve the large-scale dataset annotation issue, Apple´s Shrivastava team tried to achieve unsupervised learning of simulated images with existing computer simulation techniques and adversarial training methods, thereby avoiding the expensive image annotation process. They had three innovations, namely a ´self-regularization´ term, a local adversarial loss, and updating the discriminator using a history of refined images so that the real image is generated while retaining the input image features. The experiment results showed that the method can generate highly realistic images. The team also quantitatively analyzed the generated images by training a gaze estimation model and a hand posture estimation model. The results indicated a significant improvement over using synthetic images and achieved the state of the art on the MPIIGaze dataset without any labeled real data. However, the researchers didn´t conduct any experiment in complex scenarios involving multiple objects. The application of the proposed method still has limitations in complex scenarios.

Keywords adversarial training; unsupervised learning; simulated image; deep learning

(责任编辑 刘志远)